SAND2009-1457C

Simulation & Modeling

Arun Rodrigues

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

National Nuclear Security Administration

Sandia National Laboratories

Thanks

- IAA Simulation Working Group
- Gordon B. Bell (IBM)
- Derek Chiou (U.Texas)
- David Evensky (SNL)
- Joe Gross (U.Maryland)
- Scott Hemmert (SNL)
- Bruce Jacob (U.Maryland)
- Curtis Janssen (SNL)
- Collin McCurdy (ORNL)
- George Riley (GTech)
- Arun Rodrigues (SNL)
- Philip Roth (ORNL)
- Jeffrey Vetter (ORNL)
- Sudhakar Yalamanchili (GTech)

NNSA National Nuclear Security Administration

Sandia National Laboratories

The Problem

View of the simulation problem

Scale.....

Many Cores + Memory **X** Many Many Nodes **X** Many Many Many Threads

Multiple Audiences.....

Network Processor System **X** Application writers purchasers designers **X** system procurement algorithm co-design architecture research language research **X** present systems future systems

Complexity.....

Multi-Physics Apps Informatics Apps **X** Communication Libraries Run-Times OS Effects **X** Existing Languages New Languages

Constraints.....

Performance Cost | Power Reliability | Cooling Usability | Risk Size

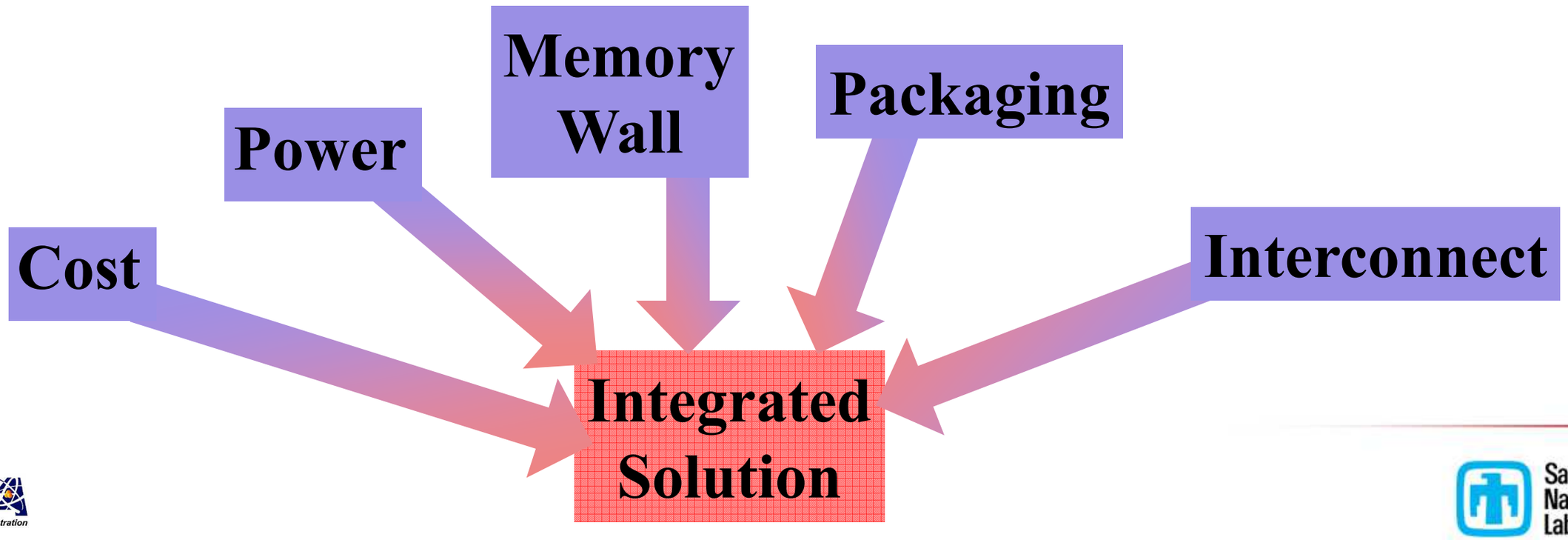
HPC Simulation: Main Challenges

- **Network/Application Feedback**: A static trace or simple statistical model will not capture the causal relationships between messages.
- **Scalability**: Many network effects only become apparent at hundreds or thousands of nodes.
- **Variable Processor/Memory/Network Systems**: Local interactions can have global performance implications.
- **Ability to Model Message Overheads:** Overheads in the network (e.g. packetization, protocol overhead) and messaging library (e.g. MPI matching, message assembly) can have a major effect on performance.
- **Ability to Explore Programming Models**: Novel hardware will require novel programming techniques and capabilities.
- **Power and Economic Effects:** Power and cost are the key limiting factors on system design. Any system model must be able provide feedback on the power and cost implications of new architectural features.

Economic & Technology

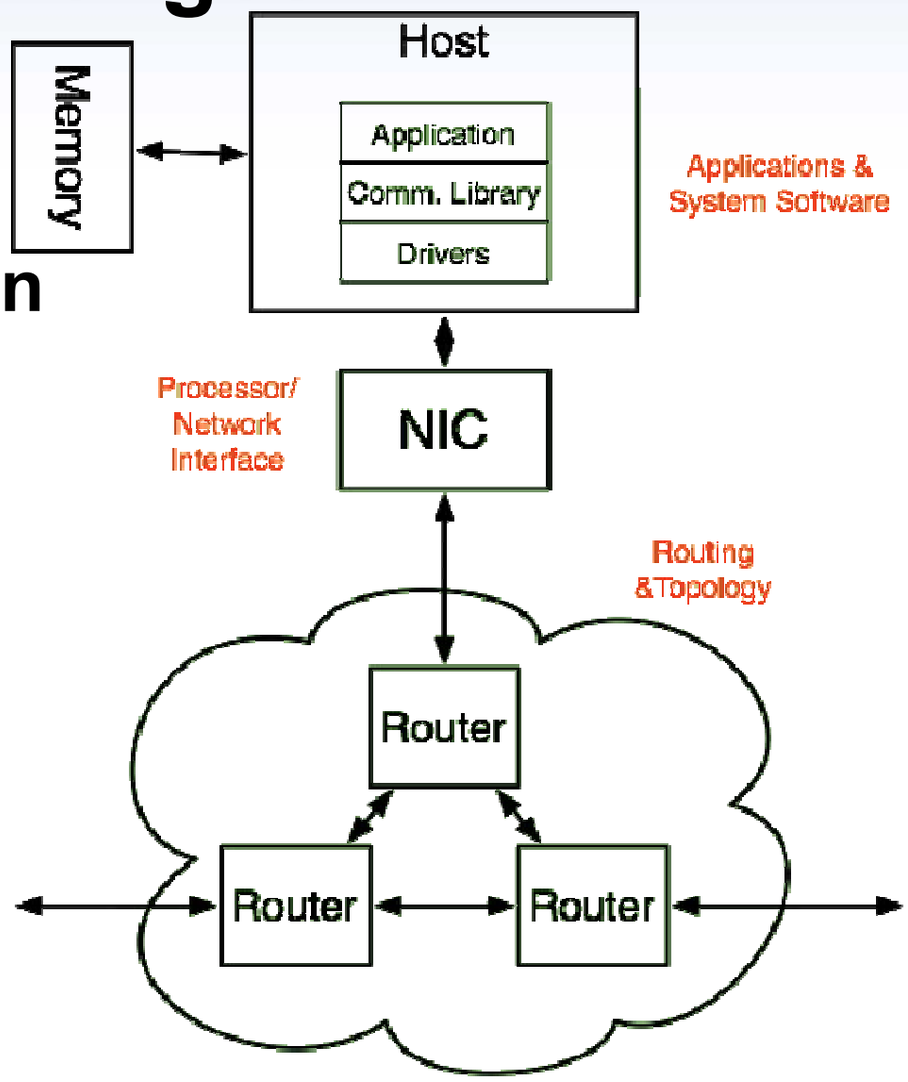
Existing industry trends not going to meet HPC application needs

- Semi-conductor industry trends
  - Moore's Law still holds, but clock speed now constrained by power and cooling limits
  - Processors are shifting to multi/many core with attendant parallelism
  - Compute nodes with added hardware accelerators are introducing additional complexity of heterogeneous architectures
  - Processor cost is increasingly driven by pins and packaging, which means the memory wall is growing in proportion to the number of cores on a processor socket
- Development of large-scale Leadership-class supercomputers from commodity computer components requires collaboration
  - Supercomputer architectures must be designed with an understanding of the applications they are intended to run
  - Harder to integrate commodity components into a large scale massively parallel supercomputer architecture that performs well on full scale real applications
  - Leadership-class supercomputers cannot be built from only commodity components



HPC Simulation Challenges

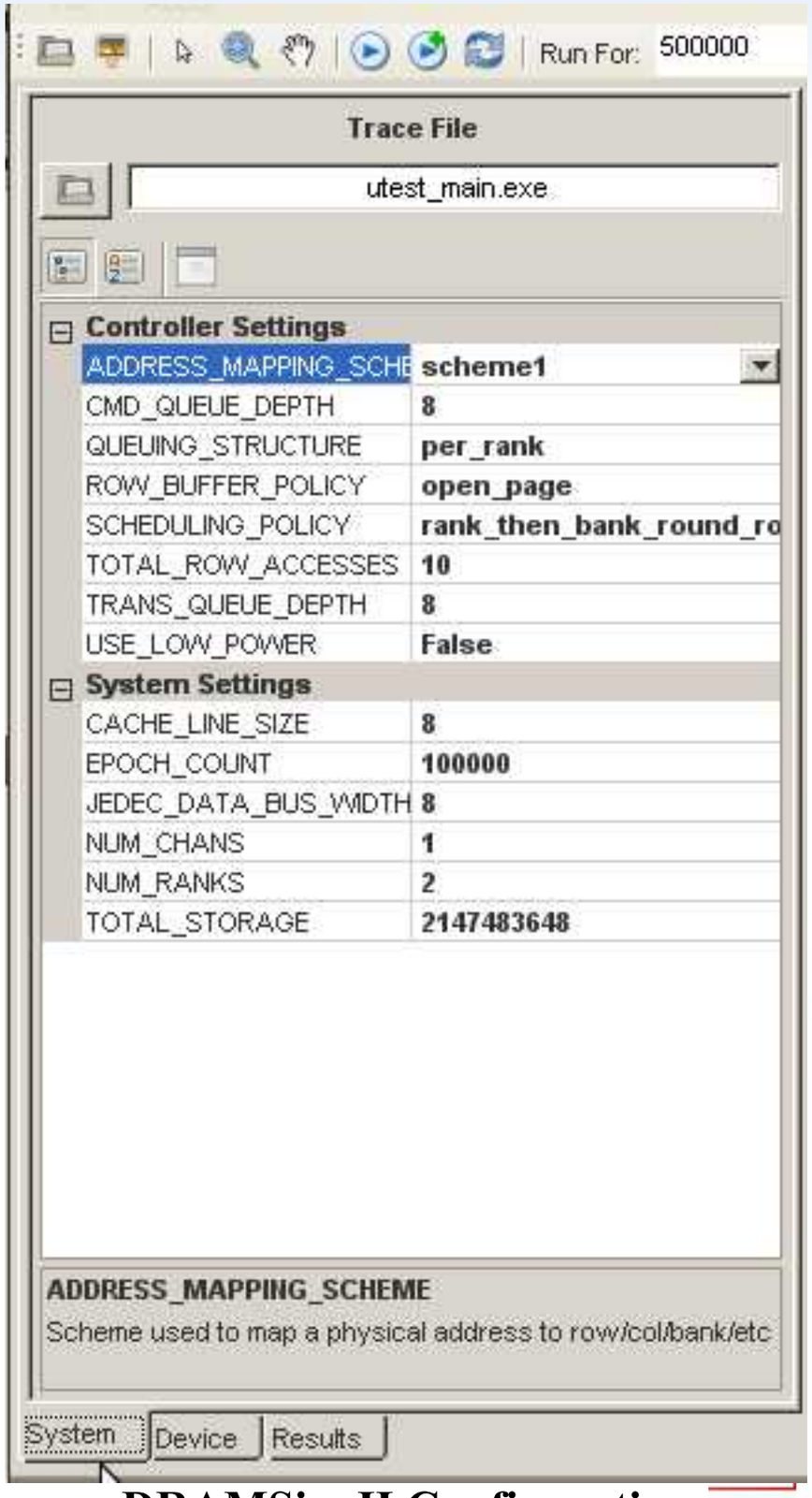
- Multiple User communities
  - Procurement: Architecture Evaluation
    - Is this the right machine to buy?
  - Application Writers: Optimization
    - How will code run on this machine?
  - Architects: Design
    - What should the machine look like?
- What level of detail?
- What subsystems examined?



- Current Simulators...
  - Cycle-accurate node-level (accurate, slow)
  - Stochastic network models (accurate?, lose details)
  - FPGA (detailed, fast, hard to develop/scale)

Challenges

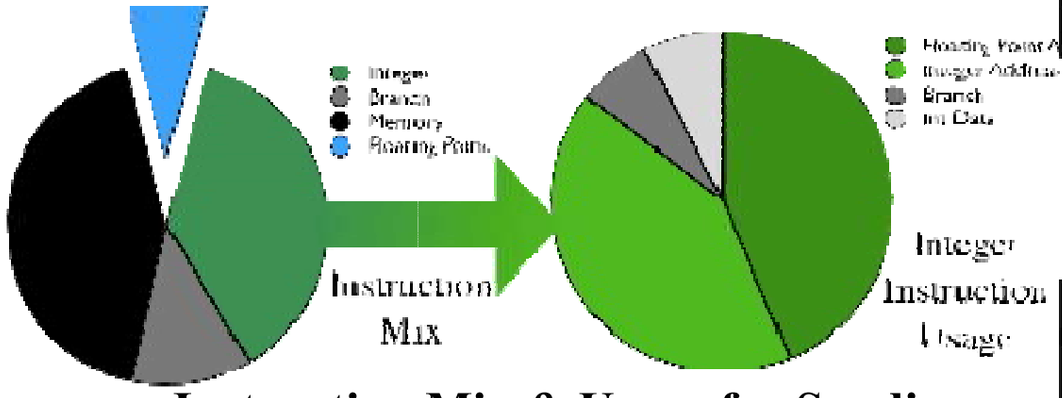
- **Usability**
  - **Porting**
  - **Support**
  - **Documentation**
  - **Configuration & Visualization**
- **Parallel execution**
  - **1000s of simulated nodes on 100s of real nodes**
  - **CAPSTONE & Gossamer Prototypes**
    - **Parallel DES simplified because of limited communication patterns**
    - **Allows use of conservative distance-based optimization**
- **Multiscale: Tradeoff between precision and speed**



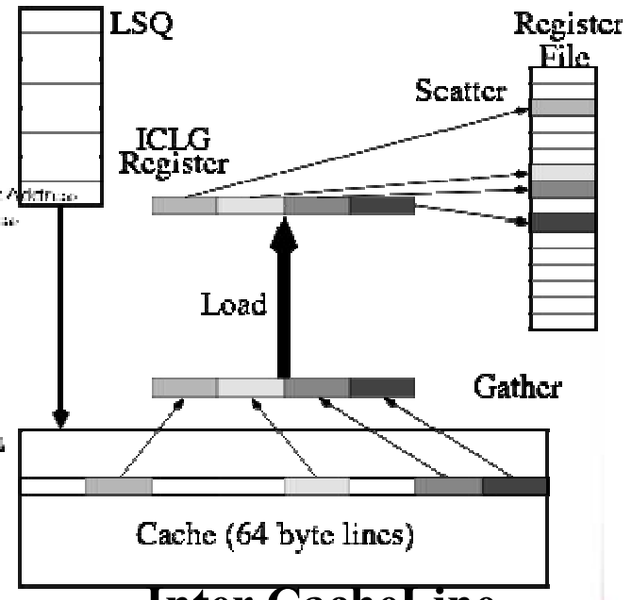
DRAMSim II Configuration Interface

Projects Supported

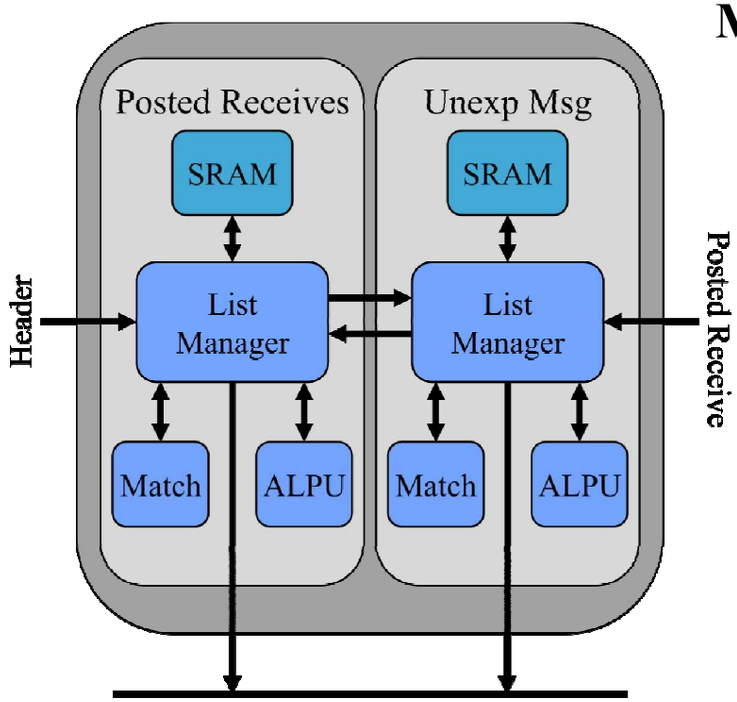
- **Microarchitecture**
  - **Inter CacheLine Gather (ICGL)**
  - **Recon. FU (Wisc./SNL)**
  - **FP Aggregates**
  - **In-Memory Ops**
- **Application analysis**
  - **Memory Footprint**
  - **Instruction Usage**
- **Network/MPI**
  - **NIC Tradeoffs**
  - **MPI Acceleration**
- **Programming Models**
  - **PIM Compiler work (SNL/Rice)**
  - **ParalIX (LSU/SNL)(FastOS)**
  - **Transactional Memory (ORNL)**
  - **QThreads**
- **Processor-In-Memory (LDRD)**



Instruction Mix & Usage for Sandia Applications



Inter CacheLine Scatter/Gather Microarchitecture



MPI Accelerator Tested with SST

Sandia National Laboratories

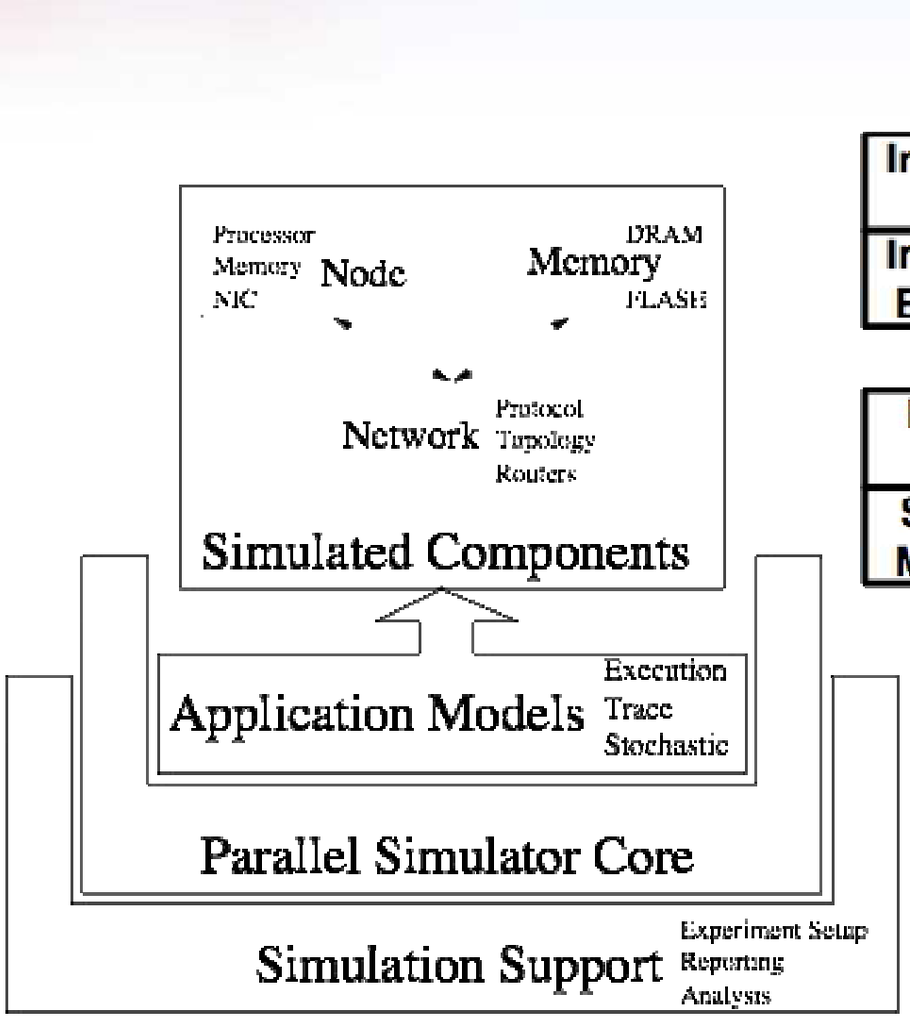
Simulator Project Goals

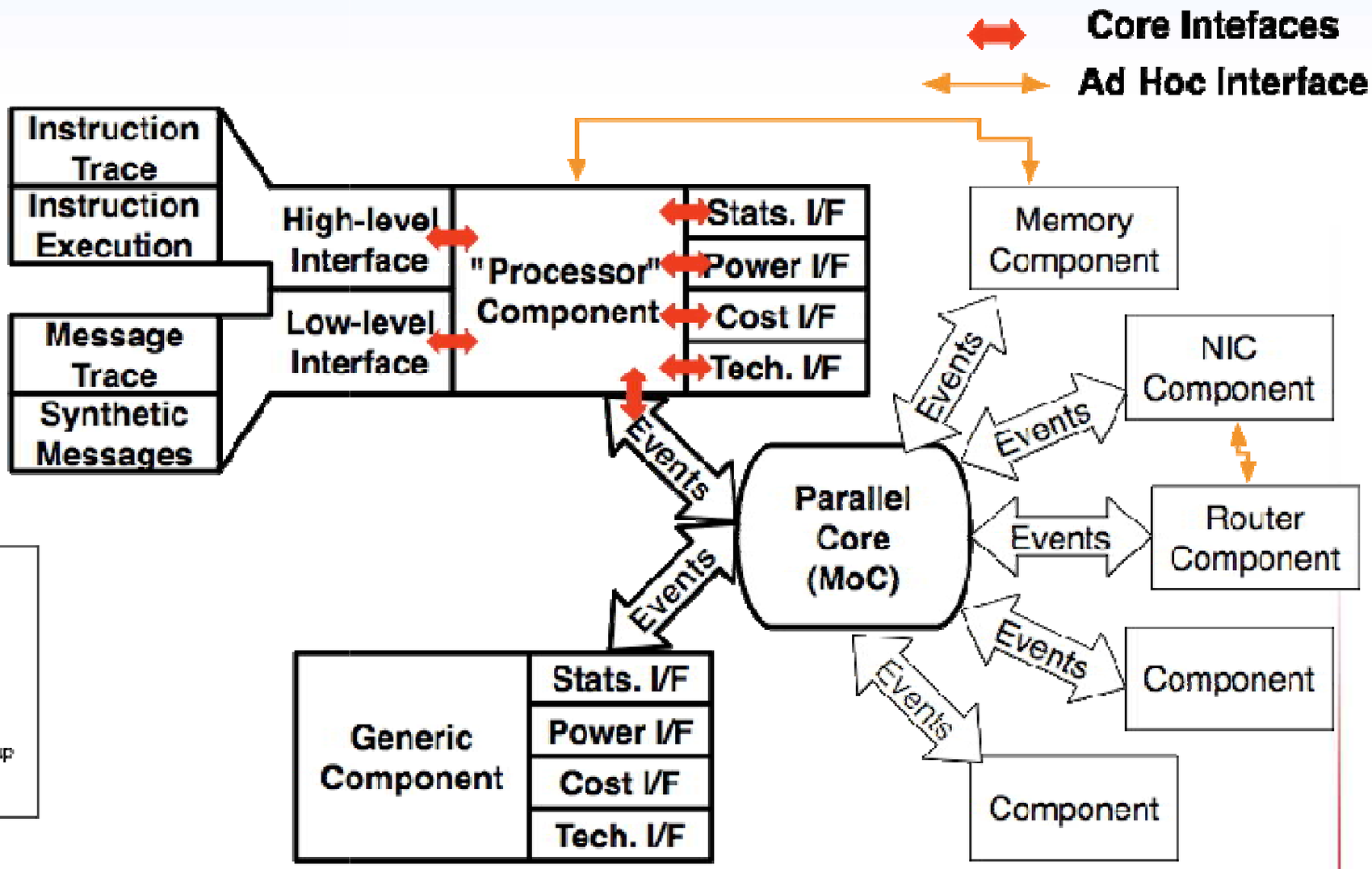
Long Term Vision	**Become <u>the</u> HPC community standard simulator**	• Long-term support • Community acceptance • Standard integration API • Advisory panel • User group
Long Term Goals	**Multi-scale simulation**	Instruction- or network-trace driven. 1:1000 to 1:100 slowdown.
	Technology model interface	Power, area, cost models
	Highly scalable parallel	10000s simulated nodes on 1000s real nodes
Near Term Goals	**Define Interfaces**	
	Prototype Parallel SST	
	Open Source Release	
	Improve Component Library	

Usage Model	Message Traces, Symbolic Workload Descriptions	Instruction Based	Execution w/ FPGA Acceleration
Real Nodes	100s-1000s	100s	1-10
Simulated Nodes	10000s-100000s	100s-1000s	1-10
Goal	System Scaling Behavior	Cycle-level system performance	Circuit level exploration, fast algorithm co-design

Expected IAA Simulator Usage Models

Proposed Architecture



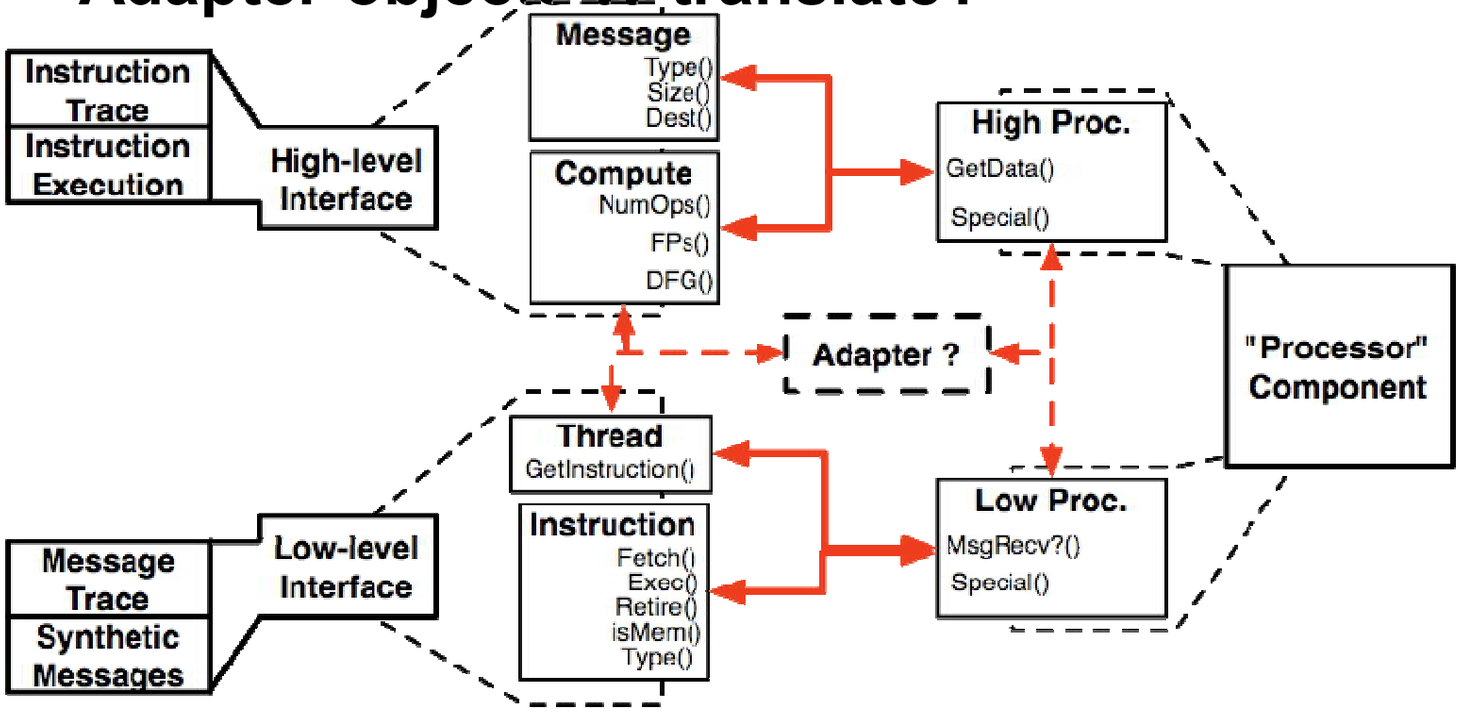


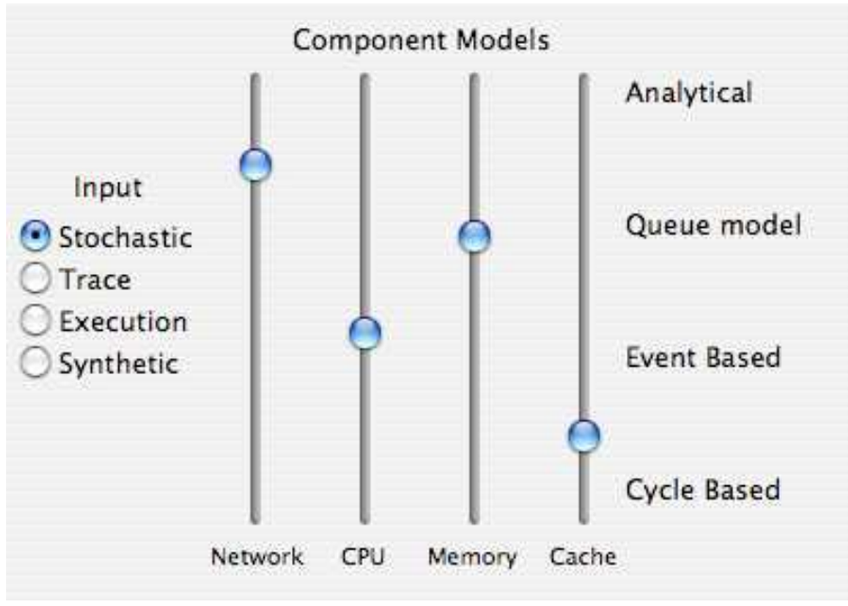
- Separate Software/Front-End from Hardware/Timing/Back-End
- Standard interfaces for power, statistics, cost?, technology?
- Event-Handling mechanism to coordinate between components

Multi-Scale Application Models

- Goal: Interface component reuse at different scales
- High- & Low-level interfaces (more?)
  - Allows multiple input types
  - Allows multiple input sources
    - Traces, stochastic, state-machines, execution...
  - Adapter objects to translate?

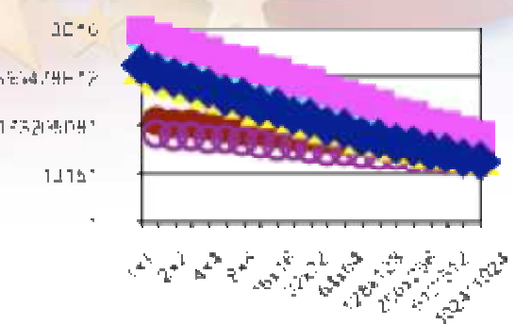
	High-Level	Low-Level
Detail	Message	Instruction
Fundamental Objects	Message, Compute block, Process	Instruction, Thread
Static Generation	MPI Traces, MA Traces	Instruction Trace
Dynamic Generation	State Machine	Execution





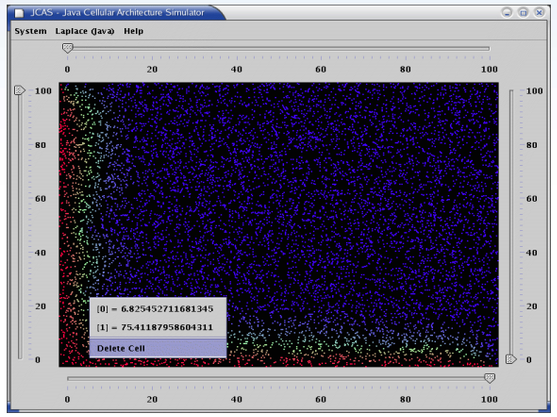
Multiscale Parameters

Heritage

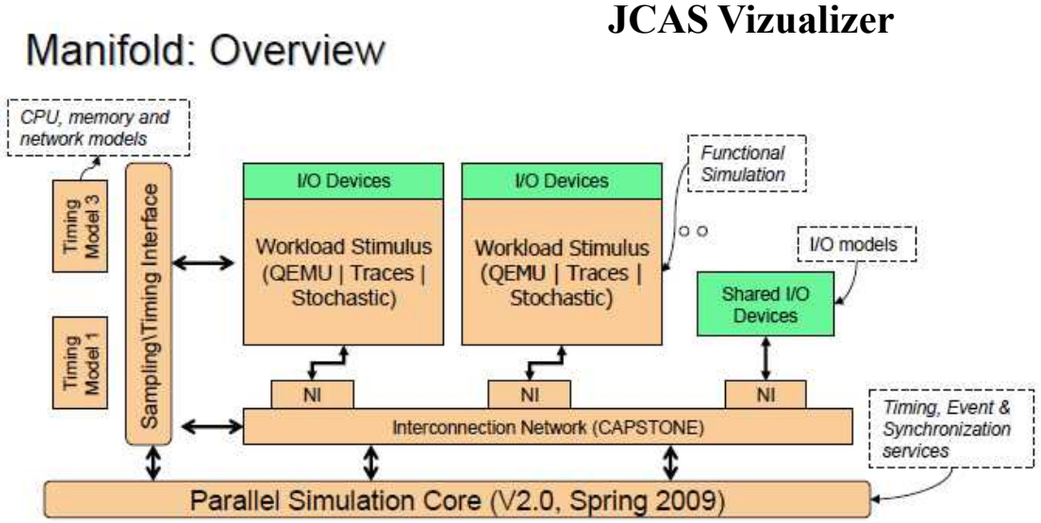


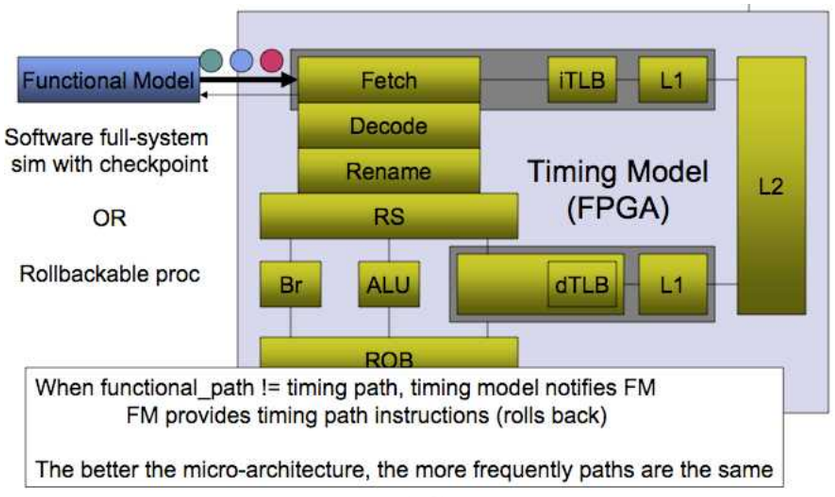
POP Grid (nproc x x nproc y cores)

Modeling Assertions



JCAS Vizualizer



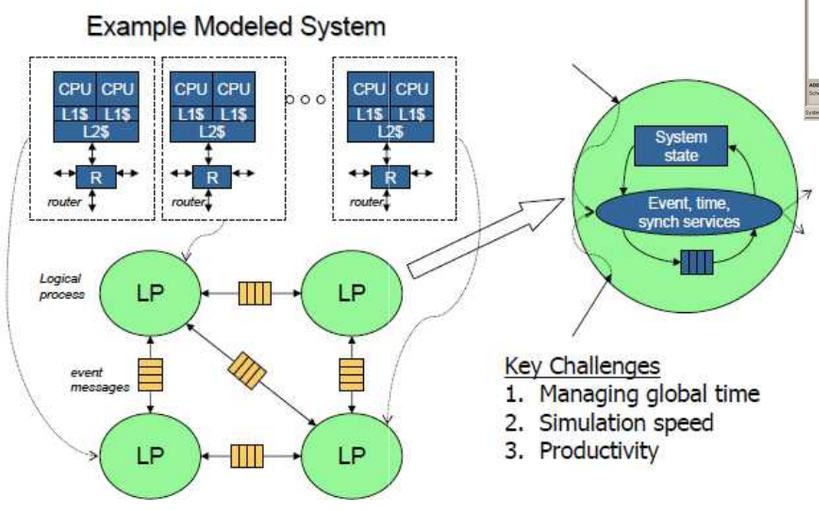


FAST

- **IAA Simulation effort is a community effort**
- **Seeking more partners...**

- **Current consortium**
  - **Sandia (Structural Simulation Toolkit)**
  - **ORNL (Scalable application models)**
  - **U. Maryland (DRAMSim II)**
  - **U.Texas-Austin (FAST)**
  - **Georgia Tech**
  - **JCAS**
  - **ArchSim**
  - **Seshat**

DRAMSim II



OAK RIDGE National Laboratory

Sandia National Laboratories

Status & Conclusions

- **IAA Simulator aims to address the key questions in HPC system simulation**
  - **Scalable simulation**
  - **Multiscale simulation**

- **IAA Simulator aims to be an effective tool for...**
  - **...system procurement**
  - **...algorithm co-design**
  - **...architecture research**

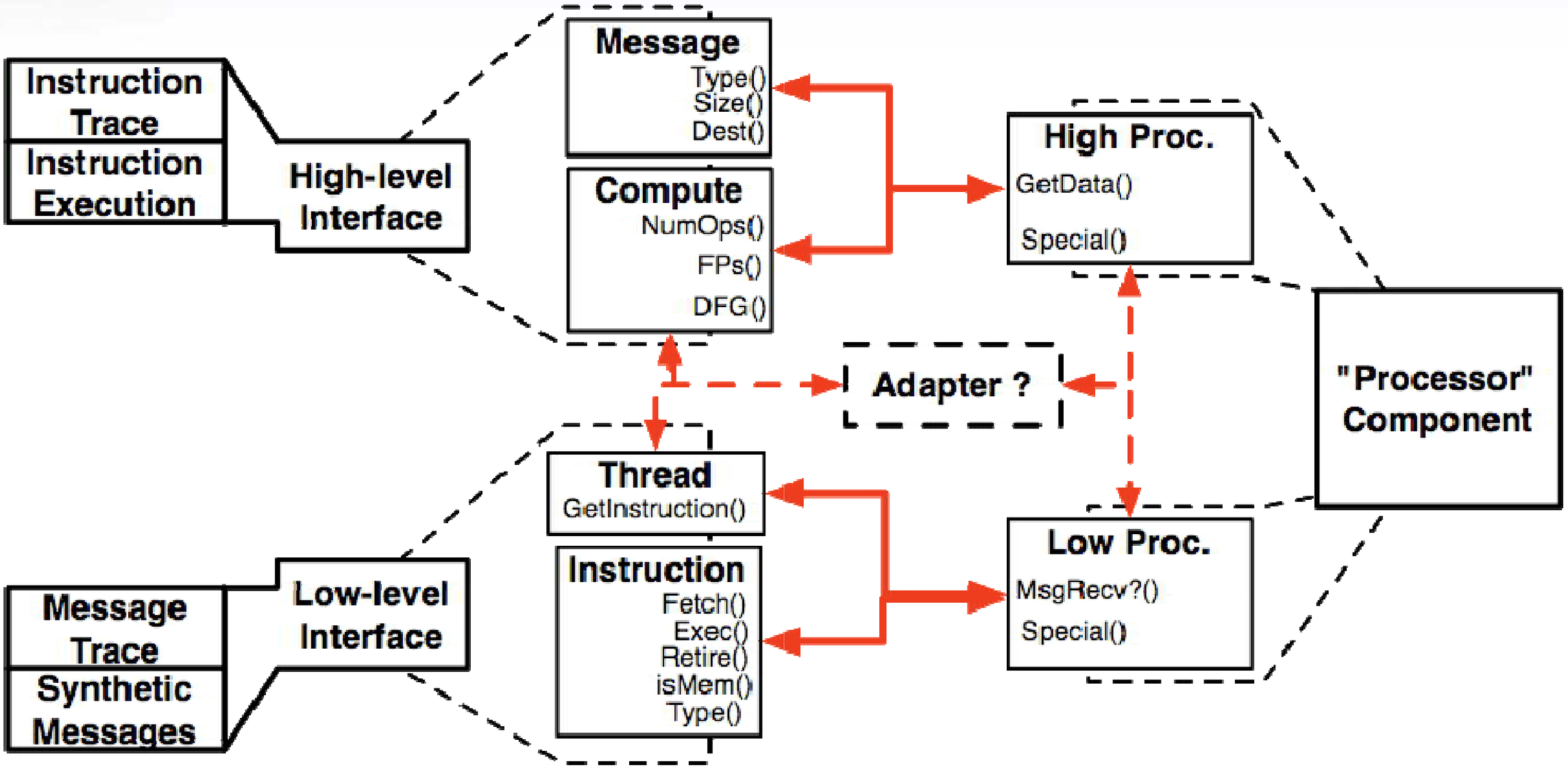
- **IAA Simulation Group is actively soliciting input from potential users and partners**
  - **What are your requirements?**
  - **What should the simulator look like?**
  - **How can we deal with your IP issues?**
  - **How would you like to be involved?**

Bonus

National Nuclear Security Administration

Sandia National Laboratories

Multiscale Front-End/Back-End



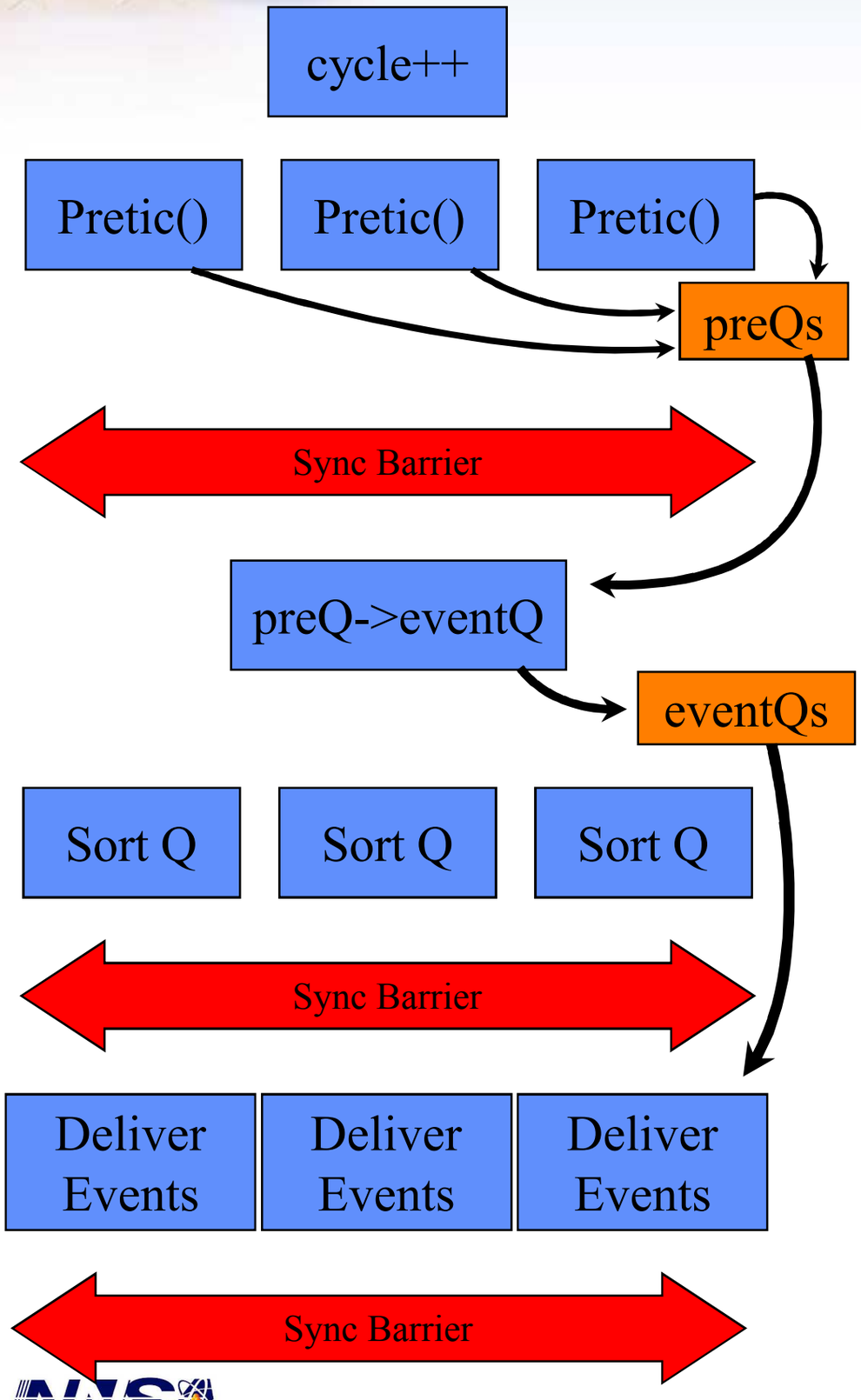
- Component reuse at different scales

Promising Technologies

- Supercomputers to design Supercomputers
- Critical mass of component simulators
  - Processors
  - Memory
  - Network
  - ...in isolation
- Critical mass of application models
  - Compact Apps
  - Scalable application models
  - State machine models
  - Message traces
- New simulation super-projects...

- SimpleScalar & PTLSim
  - ★Popular
  - – Processor oriented
  - – Conventional processor/memory model only
  - – No Network
- simg4 simg5 Tango
  - ★High accuracy
  - – Proprietary
  - – Single model processor only
- LSE/MicroLib
  - ★Very detailed
  - – Fine grained
  - – Slow execution and development time
- Simics / GEMS
  - – Not suited to network study
- Seshet - Complimentary

Run Loop



- **Parallel: Pretic()s place newly generated events into preQs**
- **Serial: preQ contents sorted into per thread eventQs**
- **Parallel: eventQs ordered**
- **Parallel: events delivered**